



**POLITECNICO**  
**MILANO 1863**

SCUOLA DI INGEGNERIA INDUSTRIALE  
E DELL'INFORMAZIONE

# Systems and Methods for Big and Unstructured Data Project

Author(s): **Gabriele Ginestroni**

**Giacomo Gumiero**

**Lorenzo Iovine**

**Nicola Landini**

**Francesco Leone**

Group Number: **10**

Academic Year: 2022-2023



# Contents

<b>Contents</b>	<b>i</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem Specification . . . . .	1
1.2 Assumptions . . . . .	1
<b>2 ER Diagram</b>	<b>3</b>
<b>3 Dataset Description</b>	<b>5</b>
3.1 Publication . . . . .	5
3.2 Author . . . . .	6
<b>4 Graph Diagram</b>	<b>9</b>
<b>5 Sample Dataset</b>	<b>11</b>
<b>6 Queries and Commands</b>	<b>13</b>
<b>7 Conclusion</b>	<b>15</b>
<b>A Appendix A</b>	<b>17</b>
<b>List of Figures</b>	<b>19</b>
<b>List of Tables</b>	<b>21</b>



# 1 | Introduction

In this chapter will be presented the problem specification and the hypothesis under which the database is implemented.

## 1.1. Problem Specification

This project aims to build an Information System that handles scientific articles contained in the DBLP bibliography. The project involves managing the type of the articles and the associated DOI (Digital Object Identifier), which identifies an article or a document and links to it on the web. Other entities to deal with are authors, identified by an ID or ORCID (Open Researcher and Contributor ID), and their affiliations with organizations. In order to address the problem, we will store data in a graph database, allowing us to visualize relations and handle information correctly.

## 1.2. Assumptions

1. All the data in the dataset are heterogeneous, so fields are different
2. The **authors** with missing field *\_id* are not considered
3. It is possible that an author writes for different organizations
4. Field *\_id* in **author** is unique
5. Field *\_id* in **article** exists and it is unique
6. It is impossible that 2 different articles are on the same journal, in the same *volume* with an intersection between *page\_start* and *page\_end*
7. The designed model doesn't take into consideration the URL associated to the article node, as the main focus of the project was not reading the article
8. It is possible to find a self-reference in a publication
9. A *venue* can be instantiated as a journal, a conference or a generic venue!!!!!!



## 2 | ER Diagram



Figure 2.1: ER Diagram Organization

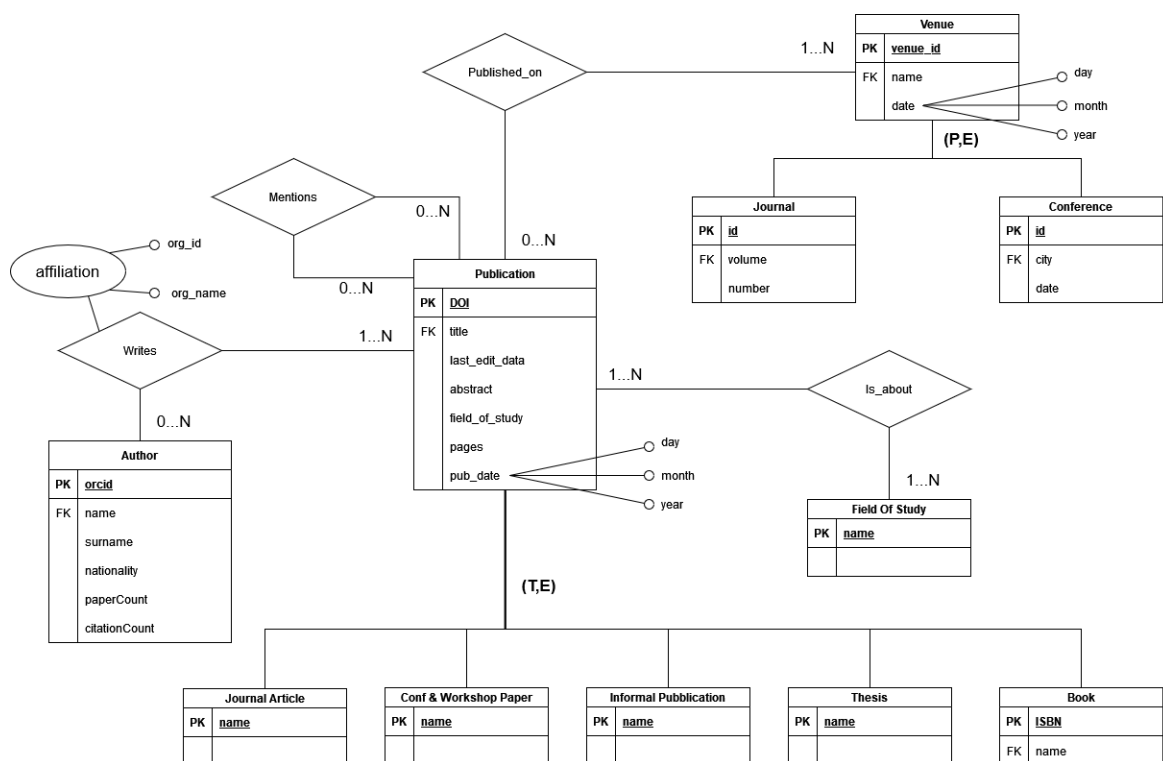


Figure 2.2: ER Diagram

The ER diagram designed contains the following entities:

- **Publication:** this entity represents all the scientific articles. They are identified by their primary key *\_id* and other important attributes are: *DOI*, *title*, *last\_edit\_data*, *abstract*, *keywords*, *pages*, *pub\_date*. Of course the attributes of such entity could be enlarged, but as a sample dataset we have believed these are enough. Publication entity is the superclass of a Total and Exclusive ISA relationship with the following subclasses: *Journal article*, *Conference & Workshop paper*, *Informal publication*, *Thesis*, *Book*
- **Author:** it represents all the people that submitted at least one publication. Its primary key is *\_id* and the foreign keys are: *name*, *surname*, *nationality*, *paper-Count*, *citationCount*. Of course the attributes of such entity could be enlarged, but as a sample dataset we have believed these are enough
- **Venue:** it's the entity that represents the type of a publication. This is a superclass that creates a Partial and Exclusive ISA relationship with the two subclasses *Journal* and *Conference*. The primary key is *raw* and the other keys are: *name*, *date*, *venue\_id*
- **Field Of Study:** this entity represents the topics of the related publication

The ER diagram designed contains the following relationships:

- **Writes:** is the relationship between *author* and *publication* which specifies also the affiliation with the *org\_id* and the *org\_name*
- **Mentions:** occurs between two *publication* and specifies when a publication refers to another one
- **Is about:** binds a *publication* with its *fields of study*
- **Published on:** simply relates a *publication* to its *venue*



## 3 | Dataset Description

In this chapter we will present all the attributes contained in the dataset pointing which of them are considered or not.

### 3.1. Publication

Publication represent the central concept of the system. In the dataset, through the profiling, we found over six thousands publication node that contains:

- **\_id** is an alphanumeric string that is the primary key, that's because is unique and every nodes owns this parameter
- **title** represents the title of the publication
- **authors** is an array of authors that will be presented later
- **venue** defines an entity that will be presented later
- **year** represents the year of publication
- **keywords** is an array containing the tag of subjects faced in the publication
- **fos** is an array containing the fields of study of the publication
- **n\_citation** is the number of times that this publication has been mentioned
- **page\_start** defines the starting page of the publication. This attribute wasn't take into consideration because doesn't target the goal of the project
- **page\_end** defines the last page of the publication. This attribute wasn't take into consideration because doesn't target the goal of the project
- **lang** represents the language of the publication
- **volume** is the volume of the publication. This attribute wasn't take into consideration due to the presence of many missing or null values

- **issue** refers to how many times a periodical has been published during that year. This attribute wasn't take into consideration due to the presence of many missing or null values
- **issn** is an identification code of a publication. This attribute wasn't take into consideration due to the presence of many missing or null values
- **issn** is an identification code of a publication. This attribute wasn't take into consideration due to the presence of many missing or null values
- **doi** Digital Object Identifier is a persistent identifier. We decided to take it into consideration due to an acceptable missing percentage, much lower than the one affecting issn or isbn attributes
- **pdf** contains a string that links to the publication PDF online. This attribute wasn't take into consideration due to the presence of many missing or null values
- **url** contains an array of links to the publication resources online. This attribute wasn't take into consideration because doesn't target the goal of the project
- **abstract** is a string containing a brief summary of the contents of the paper
- **references** is an array of ids representing the publication mentioned

### 3.2. Author

Author is the most present entity of the system with over eighteen thousands of nodes.

- **\_id** is an alphanumeric string that is the primary key, that's because is unique and almost every nodes owns this parameter
- **name** is the name of the author
- **org** is a string that represents the organization in which the author works. It is used as an attribute of the relationship *Writes* described before
- **gid** is an identifier of the author. This attribute wasn't take into consideration due to the presence of many missing or null values
- **orgid** is an identifier that represents the organization in which the author works. It is used as an attribute of the relationship *Writes* described before
- **orgs** is an array of organizations for which the author worked. This attribute wasn't take into consideration due to the presence of many missing or null values

- **email** is a string containing the email address of the author. This attribute wasn't take into consideration due to the presence of many missing or null values and because doesn't target the goal of the project
- **orcid** Open Researcher and Contributor ID is a unique identifier for authors of scientific articles. This attribute is taken into consideration although is not always present
- **oid** is an identifier for the author. This attribute wasn't take into consideration due to the presence of many missing or null values
- **bio** is a string that describes the author. This attribute wasn't take into consideration due to the presence of many missing or null values
- **sid** is an identifier for the author. This attribute wasn't take into consideration due to the presence of many missing or null values
- **org\_zh** is an identifier that represents the organization in which the author works. This attribute wasn't take into consideration due to the presence of many missing or null values



## 4 | Graph Diagram



## 5 | Sample Dataset





## 6 | Queries and Commands



## 7 | Conclusion



# A | Appendix A

If you need to include an appendix to support the research in your thesis, you can place it at the end of the manuscript. An appendix contains supplementary material (figures, tables, data, codes, mathematical proofs, surveys, ...) which supplement the main results contained in the previous chapters.



# List of Figures

2.1	ER Diagram Organization . . . . .	3
2.2	ER Diagram . . . . .	3





## List of Tables

